BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59178
doi: 10.3897/biss.4.59178

OPEN ACCESS

Conference Abstract

The Standards behind the Scenes: Explaining data from the Plazi workflow

Donat Agosti[‡], Marcus Guidoti[§], Terry Catapano[‡], Alexandros Ioannidis-Pantopikos[|], Guido Sautter[¶]

‡ Plazi, Bern, Switzerland
§ Plazi, Porto Alegre, Brazil
| CERN, Meyrin, Switzerland
¶ IPD Böhm, Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author: Donat Agosti (agosti@amnh.org)

Received: 30 Sep 2020 | Published: 09 Oct 2020

Citation: Agosti D, Guidoti M, Catapano T, Ioannidis-Pantopikos A, Sautter G (2020) The Standards behind the Scenes: Explaining data from the Plazi workflow. Biodiversity Information Science and Standards 4: e59178. https://doi.org/10.3897/biss.4.59178

Abstract

As part of the CETAF COVID19 task force, Plazi liberated taxonomic treatments, figures, observation records, biotic interactions, taxonomic names, and collection and specimen codes involving bats and viruses from scholarly publications with the intention to create open access, findable, accessible, interoperable and reusable data (FAIR). The data is accessible via TreatmentBank and the Biodiversity Literature Repository (BLR) and it is continually harvested and reused by the Global Biodiversity Information Facility (GBIF) and Global Biotic Interactions (GloBI). This data was processed, enhanced and liberated by the Plazi workflow, which involves a dedicated infrastructure including a desktop application (GoldenGate Imagine) that converts portable document format files (PDF) to a dedicated open compressed file format (Image Markup File (IMF)) that is responsible for the data enhancement.

To enhance the data contained in the publications, including the biological interactions, a series of standards and vocabularies are used. To the exception of TaxPub, which is a taxonomic specific extension of the U.S. National Center for Biotechnology Information's (NCBI) Journal Article Tag Suite (JATS), all other used vocabulary were previously proposed. This goes along with Plazi's mission to reuse standards unless they are not available. The following standards of vocabularies are used: Metadata Object Description

Schema (MODS) to model article metadata information on Plazi's XMLs; Darwin Core for taxonomic ranks and materials citation related data; Open Biological and Biomedical Ontology (OBO); Relations Ontology for biological interactions between organisms. The latter two are also used in the custom metadata in the Biodiversity Literature Repository at Zenodo.

In this presentation we will provide an overview of the different types of data followed by the standards or vocabularies applied for every and each one of them and their parts. The goal is to provide the context on how the data liberated by Plazi is described, which is extensively reused by third-party applications such as GBIF or GloBI. The use of the standards allows fully automated, daily data ingests by GBIF.

Keywords

biodiversity, data conversion, standards, MODS, TaxPub, JATS, Darwin Core, OBO, biotic interactions

Presenting author

Marcus Guidoti

Presented at

TDWG 2020